Nonparametric Hierarchical Modeling for Detecting Boundaries in Areally Referenced Spatial Datasets

Pei Li*, Sudipto Banerjee, Timothy E. Hanson and Alexander M. McBean

* Division of Biostatistics, University of Minnesota,
420 Delaware Street SE, MMC-303,
Minneapolis, MN 55455.
Ph: 612.624.4655;
Fax: 612.626.0660;
email: lixx0525@umn.edu.
Nonparametric Hierarchical Modeling for Detecting Boundaries in Areally Referenced Spatial Datasets

Abstract

With increasing accessibility to Geographical Information Systems (GIS) software, researchers and administrators in public health are increasingly encountering spatially referenced datasets. Inferential interest of spatial data analysis often resides not in the statistically estimated maps themselves, but on the formal identification of “edges” or “boundaries” on the map. Boundaries can be thought of as a set of connected spatial locations that separate areas with different characteristics. A class of nonparametric bayesian models are proposed in this paper to account for uncertainty at various levels to elicit spatial zones of rapid change that suggest hidden risk factors driving these disparities. Simulation study are conducted to illustrate the new approaches and compare with existing methods. “Boundaries” on Pneumonia and Influenza hospitalization map from the SEER-Medicare program in Minnesota are detected using the proposed approaches.

Keywords: Areally referenced spatial data; Difference boundary; Conditional autoregressive model; Dirichlet process; Stick-Breaking process.
1 Introduction

Geographical Information Systems (GIS) software has revolutionized the analysis of spatially referenced health data with its depiction of counts and rates over study areas. In public health services, to protect patient privacy, spatial data are usually available as case counts or rates aggregated over areal regions (e.g., counties, census-tracts or ZIP codes) rather than the geographical location of the individual residences. When spatial dependence in the data renders the ordinary least squares regression model unsuitable, alternative models that incorporate spatial dependence should be considered. For areally-referenced data, the association structures are built upon adjacencies or neighborhood structures of the regions. Here the statistical models regard observations from a region to be more similar to those from its neighboring regions than those arising from regions farther away. Two such models are the spatial lag model and the spatial error model. They have been widely employed for smoothing maps and evincing spatial trends and clusters in econometrics (e.g., Anselin, 1988; Le Sage and Pace, 2009) and in public health (e.g., Banerjee et al., 2004).

The models for areal data are primarily used for smoothing maps revealing spatial trends and identifying clusters. Subsequent inferential interest often resides in the formal identification of “barriers” or “boundaries” on the spatial surface or map. The ‘boundary’ here is a border with sharp changes in outcome on either side. A ripe area of research is the statistical detection of spatial or geographical barriers (also known as difference boundaries) that cause major differences in outcomes between neighboring areal units. Statistical models can help analysts separate significant boundaries from those arising out of random noise in the data. This ‘boundary’ detection problem is often referred to as “wombling”, after a foundational article by Womble (1951). While statistical boundary analysis has been applied extensively to point-referenced and gridded (or lattice) data (see, e.g., Banerjee and Gelfand, 2006), formal statistical inference in areal contexts present unique challenges that we outline
later. Our proposed framework intends to overcome these challenges using a novel class of probability distributions.

Algorithmic approaches to areal wombling, also known as polygonal wombling, have been addressed by Jacquez and Greiling (2003a, 2003b). While attractive in their simplicity and ease of use, the algorithmic approaches do not account for all sources of uncertainty and can lead to spurious statistical inference. For instance, it is well-known that in public health contexts extremeness in counts and rates corresponding to certain thinly populated regions arises due to random variation in the observed data rather than any systematic differences. As a remedy, Lu and Carlin (2004), Lu et al. (2007) and Ma, Carlin and Banerjee (2008) investigated estimating the adjacency matrix within a hierarchical framework using priors on the edges. However, these models often involve weakly identifiable parameters that are difficult to estimate without informative prior knowledge that is usually unavailable.

Our primary contribution is a more flexible class of spatial models that will not only deliver better inference for areally aggregated health outcome data, but will also provide stochastic assessments regarding the presence of geographical barriers. We achieve this through classes of more flexible and robust non-parametric Bayesian hierarchical models. Section 2 offers a brief exposition to models for areally referenced count data. Section 3 elucidates the key issues in areal boundary analysis and our Bayesian nonparametric modeling approaches. Sections 4 and 5 discuss a simulation study and the analysis of a Minnesota Pneumonia & Influenza (P & I) dataset to detect “spatial health barriers” between neighboring counties in Minnesota. Finally, Section 6 concludes the article with an eye towards future work.
2 Hierarchical models for areal data

Most of the data types alluded to in the preceding section can be analyzed using Bayesian hierarchical models that incorporate geographical effects. To illustrate such models with count data, let $Y_i$ be the observed number of patients who underwent a specific preventive or clinical outcome in areal unit $i$, $i = 1, \ldots, n$, and let $E_i$ be the expected number of outcomes for that unit. We treat the $Y_i$ as random variables, while the $E_i$ are assumed fixed and known (and are often simply taken as the expected number of outcomes in the region assuming the outcome is equally likely across space). For rare outcomes, a Poisson model approximation to a binomial sampling distribution for outcome counts is often used. Thus, a commonly used likelihood when mapping a single outcome is

$$Y_i \sim^{ind} \text{Poisson}(E_i e^{\mu_i}), \ i = 1, \ldots, n,$$ (1)

where $\mu_i = x_i' \beta + \phi_i$ represents the log-relative risk, estimates of which are often based on the departures of observed from expected counts. The $x_i$ will include explanatory, region-level covariates or predictors, having parameter coefficients $\beta$. Bayesian inference will deliver exact inference through the posterior distribution of $\beta$ conditional upon the data.

Each $\phi_i$ represents the spatial random effect associated with region $i$. A popular approach builds a joint distribution on the spatial random effects using conditional distributions. More precisely, we specify, for each $\phi_i$,

$$\phi_i \mid \phi_j, j \neq i, \sim N \left( \frac{\rho}{m_i} \sum_{i \sim j} \phi_j, \frac{1}{\tau m_i} \right), \ i, j = 1, \ldots, n,$$ (2)

where $i \sim j$ denotes that region $j$ is a neighbor (typically defined in terms of spatial adjacency) of region $i$. Using the theory of Markov random fields (e.g. Cressie, 1993; Banerjee
et al., 2004, Ch.3), it can be shown that (2) implies the following joint distribution for the spatial effects:

$\phi \sim N_n(0, \sigma^2(D - \rho W)^{-1})$ ,

(3)

where $N_n$ denotes the $n$-dimensional normal distribution, $D$ is a $n \times n$ diagonal matrix with diagonal elements $m_i$ that denote the number of neighbors of area $i$, and $W = \{w_{ij}\}$ is the adjacency matrix of the map (i.e., $w_{ii} = 0$, and $w_{ij} = 1$ if $i$ is adjacent to $j$ and 0 otherwise). In the joint distribution (3), $\sigma^2$ is the spatial dispersion parameter, and $\rho$ is a spatial autocorrelation parameter. We denote this distribution concisely as $CAR(\rho, \sigma^2)$.

One needs to ensure that $D - \rho W$ is positive definite, a sufficient condition for which (e.g., Banerjee et al., 2004) is to restrict $\rho \in (1/\lambda_{(1)}, 1)$, where $\lambda_{(1)}$ is the minimum eigenvalue of the adjacency matrix $W$ (it can be shown that $\lambda_{(1)} < 0$; see, e.g., Banerjee et al., 2003).

The CAR model has been especially popular in Bayesian inference as its conditional specification is convenient for Gibbs sampling and MCMC schemes. The CAR structure (3) reduces to the well-known intrinsic conditionally autoregressive (ICAR) model if $\rho = 1$, or an independence model if $\rho = 0$. The ICAR model induces “local” smoothing by borrowing strength from the neighbors, while the independence model assumes independence of spatial rates and induces “global” smoothing. The smoothing parameter $\rho$ in the CAR prior (3) controls the strength of spatial dependence among regions, though it has long been appreciated that a fairly large $\rho$ may be required to deliver significant spatial correlation.
3 Bayesian nonparametric models for areal data

3.1 Modelling considerations for areal boundary analysis

Formal statistical methods for areal boundary analysis can be approached from a few different perspectives. One approach would regard it as a huge model comparison problem, where each model represents a different boundary hypothesis. A boundary hypotheses corresponds to a particular underlying map specifying which edges should be smoothed over and which should not. A few issues, however, arise regarding the exact choice of the model. For instance, consider the hypothesis of no difference boundaries at all in the map. What model would correspond to this hypothesis? If we believe there are no difference boundaries at all, should we consider the map as comprising essentially a single homogeneous region? The latter implies having no region-specific effects at all and would reduce the analysis to one of simple regression; hence no random effects in (1). Alternatively, we could still regard $Y_i$ as arising from $n$ different regions but, given the absence of difference boundaries, we would retain independent regional effects instead of spatial structures in the model. This yields a linear random effects model with iid regional effects. The choice is not straightforward and would depend upon the objectives of the analysis.

In the second approach, we always retain random effects and consider models varying in their specification of the neighborhood matrix $W$ that controls spatial smoothing. At the other extreme all the geographical edges may in fact be difference boundaries. Any intermediate model that lies between these extremes is completely specified by modifying the original map to delete some edges. Ideally we would like to consider a class of models $\mathcal{M} = \{M_1, \ldots, M_K\}$ representing all possible models or all possible maps derived from $W$ by deleting combinations of geographical edges. In other words, $M_k$ denotes a model with the adjacency matrix $W_k$ that has been derived by changing some of the 1’s to 0’s in $W$. 
This amounts to dropping some edges from the original map or, equivalently, combining two regions into one. However, now we encounter an explosion in the number of models. To be precise, if $W$ is the original geographical map, we have $2^{|W|/2}$ models to compare. This will require sophisticated MCMC Model Composition or MC$^3$ algorithms for selecting models. These methods will become computationally intensive and in conducive to learning about edge effects in relatively large maps.

A different approach seeks to estimate the adjacency matrix within a hierarchical framework using priors on the adjacency relationships. These involve incorporating “edge effects”, i.e. random effects corresponding to the edges, in addition to regional effects. These edge effects would be modelled by another CAR model, or some other MRF, leading to rather complex site-edge models (Ma et al., 2008). However, these models often involve weakly identifiable parameters that are difficult to estimate from the data. Fairly informative prior knowledge is required that is usually unavailable causing the MCMC algorithms to be inexorably slow in converging to the desired posterior distributions.

Instead of incorporating random “edge effects”, we propose to explore an alternative stochastic mechanism that would let us detect difference boundaries by considering probabilities such as $P(\phi_i = \phi_j | i \sim j)$. Clearly using direct CAR specifications will simply not work as that yields continuous measures for the $\phi_i$’s, rendering $P(\phi_i = \phi_j | i \sim j) = 0$. The challenge here is to model the spatial effects in an almost surely discrete fashion while at the same time accounting for the spatial dependence. A nonparametric Bayesian framework that models the spatial effects as almost surely discrete realizations of some distribution comes to mind – the Dirichlet process (Ferguson, 1973) presents itself as a natural choice, but how do we accommodate spatial (areal) dependence? This is a novel methodological problem and we addresses this issue in the subsequent sections.
3.2 Dirichlet Process Mixture (DPM) models for clustered data

The Dirichlet Process Mixture (DPM) model is a very popular choice for analyzing clustered data. This model encourages clustering of observations without borrowing information from geographical neighbors. We briefly describe this below. In the context of (1), the DPM model can be written as

\[
Y_i | \beta, \phi_i \sim \text{Poi} \left( e^{x_i' \beta + \phi_i} \right); \quad \phi_i | G \sim G(\theta); \\
G \sim \text{DP}(\alpha, G_0(\theta)).
\]  

(4)

The Dirichlet process \( DP(\alpha, G_0) \) is defined as the distribution of a probability measure \( G \) over a measurable space \( (\Omega, B) \) such that, the random vector \( (G(A_1), ..., G(A_k)) \) is distributed as a finite-dimensional Dirichlet distribution with parameters \( (\alpha G_0(A_1), ..., \alpha G_0(A_k)) \) for any finite measurable partition \( (A_1, A_2, ..., A_k) \) of \( \Omega \). Here \( \alpha \) is a real-valued parameter and \( G_0 \) is a probability distribution called the base measure. The Dirichlet distribution is a family of continuous multivariate probability distributions parameterized by a vector of positive real numbers. It is the multivariate generalization of the beta distribution and forms a conjugate prior for the multinomial distribution in Bayesian statistics. Further technical details on DP’s are available in Ferguson (1973) from a formal probabilistic perspective.

When the \( \phi_i \)'s are modelled as in (4), then marginalizing over the measure \( G \) yields the predictive distribution of \( \phi_{n+1} \), given \( \{\phi_i\}_{i=1}^n \) as

\[
\phi_{n+1} | \{\phi_i\}_{i=1}^n, \alpha, G_0(\cdot) = \frac{\alpha G_0(\cdot) + \sum_{i=1}^n \delta_{\phi_i}(\cdot)}{\alpha + n}.
\]  

(5)

Thus the number of distinct values of \( \phi_j \)'s is controlled by the precision parameter \( \alpha \). For \( i \geq 1 \), the observation \( \phi_i \) takes on a new value with probability \( \alpha/(\alpha + i - 1) \), thus the
expected number of distinct values of $\phi$ is $\sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1}$. The conditional distributions in (5) can be used to determine the joint distribution of $\phi_i$'s conditional upon $G_0$ and $\alpha$, i.e. after integrating out $G$. Blackwell and MacQueen (1973) related the DP to a generalized *Polya urn scheme* that leads to effective sampling strategies for the spatial effects.

We describe a class of hierarchical nonparametric Bayesian models for areal data that would facilitate boundary detection. A key ingredient for these models is the ‘stick-breaking” representation Dirichlet process (DP) proposed by (Sethuraman, 1994). The “stick-breaking” representation says that a draw $G$ from the Dirichlet process can be written as $G(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(\cdot)$, where $\delta_{\theta_i}$ is the Dirac measure (point mass) located at $\theta_i$, each $\theta_i$ is a random draw from the base distribution $G_0$, $p_i = V_i \prod_{l=1}^{i-1} (1 - V_l)$ with $p_1 = V_1$, where each $V_i \overset{iid}{\sim} Beta(1, \alpha)$. The $p_i$'s are called the “stick-breaking” weights (their infinite sum equals 1) and the $\theta_i$'s are called atoms. In practice, it is reasonable to replace the infinite sum with the sum of the first $N$ terms, since the probability mass in each term decays rapidly. We can simply let $V_N = 1$ to truncate the sum to finite terms.

The “stick-breaking” representation offers an extremely rich framework for constructing flexible Bayesian nonparametric models of which the DP’s and DDP’s constitute but a small subclass. In fact, for every conceivable joint distribution on the stick-breaking weights and the atoms, there is associated a stick-breaking stochastic process. One extension yields Dependent Dirichlet processes (DDP) that attempt to introduce dependencies through the stick-breaking weights and the atoms. Other DDP models proposed by recent authors include kernel stick-breaking processes by Dunson and Park (2008) (see also Dunson, Pillai and Park, 2007) and a nested Dirichlet process for modeling a collection of dependent distributions (Rodriguez, Dunson and Gelfand, 2006). Nonparametric Bayesian models for spatial datasets have also recently been receiving much attention. Much of this work has focused upon a class of spatial Dirichlet process (SDP) mixture models introduced by Gelfand, Kottas and
MacEachern (2005). The SDP is a Dirichlet process (DP) defined on a space of surfaces that yields almost surely discrete realizations with countable support. Duan, Guindani and Gelfand (2007) extended the SDP by allowing different surface selection at different sites. Reich and Fuentes (2007) develop a spatial stick-breaking prior (SSB) to analyze hurricane surface wind fields.

The aforementioned work apply exclusively to point-referenced data and not to areal data. Our need to move into DP’s (and their extensions) is even more fundamental – accommodating non-zero probability masses for spatial random effects without sacrificing richness in any other way for areal models is problematic. The models we propose below correspond to a subclass of stick-breaking process priors that includes the DP and the SDP as special cases. In particular, we construct an areally-weighted stick-breaking process (ARSB) and an areally-referenced Dirichlet Process (ARDP) that will act as versatile models for areal data allowing formal boundary analysis. These are also easily adapted to multivariate settings using multivariate SAR or CAR models.

3.3 Areally-Referenced Spatial Stick-Breaking Prior

Existing dependent DP models offer Bayesian nonparametric modelling frameworks applicable to continuous covariates. The method adapts easily to point-referenced settings taking the coordinates of the locations as the covariates. On the other hand, for areally-referenced spatial data, the underlying spatial association is built on adjacency or neighborhood structures of the regions, hence the covariates related to the spatial locations are not continuous everywhere. DDP’s allowing continuous predictor-dependent weights no longer apply.

In lieu of the above, here we propose an areally-referenced stick-breaking prior (ARSB) for the spatial random effects that will apply to data collected on areal units. We adapt a point-referenced spatial stick-breaking approach of Reich and Fuentes (2007) to areal data
by incorporating spatial dependence in the DP by further weighing the proportion of the stick broken each time during the stick-breaking process. These additional weights are then allowed to borrow strength across the neighbors by specifying their joint distribution to be a CAR (Sec. 2). Here we discuss univariate outcomes; multivariate extensions are possible using multivariate CAR (MCAR) distributions (Banerjee et al., 2004).

The spatial random effects are assigned a stick-breaking prior, whose weight parameters

\[ p_{ik} = w_{ik} V_k, \quad i = 1, \ldots, n, \quad k = 1, 2, \ldots, \]

depend not only on the \( V_k \)'s, but also on “location” weight parameters \( w_{ik} \). These weights lie between \((0, 1)\) so that each \( p_{ik} \) is a valid stick-breaking weight \((0 < p_{ik} < 1)\). Since the CAR distribution has support over the entire real line, we introduce a transformation \( \logit(w_{ik}) = z_{ik} \) and allow the \( z_{ik} \)'s to be distributed as a CAR. Of course, any other link mapping the unit interval to the real line could have been used, e.g. the probit (we intend to explore these choices further). For each \( k \), we let \( \{z_{ik}\}_{i=1}^n \) be distributed as a CAR distribution yielding a Markov random field (MRF) on the location weights and allows the desired smoothing across neighbors. Maximum smoothing is obtained by setting \( \rho = 1 \) which yields the well-known intrinsic CAR (ICAR) prior. This prior is improper as \( D - W \) is singular, but for a map without islands this issue can be resolved by imposing the additional constraint \( \sum_{i=1}^n z_{ik} = 0 \).

The ARSB model, truncated to \( m \) terms for the stick-breaking representation, with a first stage Poisson likelihood is:

\[
Y_i \sim \text{Poi}(e^\mu_i); \quad \mu_i = x_i' \beta + \phi_i,
\]

\[ \phi_i \sim G^{(i)}; \quad G^{(i)}(\cdot) = \sum_{k=1}^m p_{ik} \delta_{\theta_k}(\cdot), \quad \theta_k \sim N(0, \sigma^2_k) \]

\[ p_{i1} = w_{i1} V_1, \quad p_{ik} = w_{ik} V_k \prod_{l=1}^{k-1} (1 - w_{il} V_l), \quad V_k \overset{iid}{\sim} \text{Beta}(1, \alpha), \]

\[ \{z_{ik}\} \sim \text{CAR}(\rho, \sigma^2_k), \quad k = 1, \ldots, m \]
where \( \logit\{z_{ik}\} = w_{ik} \). Recall that the \( \alpha \) parameter controls the number of distinct values among the \( n \) observations. The covariance between dependent variables \( Y_i \) and \( Y_j \) is induced by the covariance of the spatial random effects \( \phi_i \) and \( \phi_j \). We also let \( \theta_k \overset{iid}{\sim} N(0, \tau^2) \).

### 3.4 Areaty-referenced Dirichlet process

The ARSB model we introduced in the last section is a very flexible spatial model allowing the dependence between the discrete distributions on different regions. However, this model does not yield tractable marginal distributions of the \( \phi_i \). Duan, Guindani and Gelfand (2007) introduced random distributions for the spatial effects associated with point-referenced data allowing different surface selection at different sites. They specified the model so that the marginal distribution of the effect at each site still comes from a Dirichlet process. However, their approach requires replications and is suited for point-referenced data. Here we propose an areal alternative, which we call an areally-referenced Dirichlet process (ARDP). The ARDP maintains the marginal distribution of each spatial random effect to be a regular univariate DP while incorporating the spatial dependence between these DPs.

Consider spatial random effects \( \phi_i, i = 1, \ldots, n \) arising from an identical random measure \( G \), where \( G \sim DP(\alpha, G_0) \). We introduce spatial dependence between these DPs by constructing dependent uniform \((0,1)\) random variables. Suppose \( \gamma_1, \ldots, \gamma_n \) are jointly distributed as a \( CAR(\rho) \), and \( F^{(1)}(\cdot), \ldots, F^{(n)}(\cdot) \) denote the cumulative distribution functions of the marginal distribution of the components of the CAR random variable. Marginally, each \( F^{(i)}(\gamma_i) \) is uniform \((0,1)\) but they will be dependent through \( \gamma_1, \ldots, \gamma_n \).

More explicitly, we formulate our hierarchical areally-referenced Dirichlet process (ARDP) model as follows. We use a Poisson likelihood for the first stage model, but this could be
replaced by any member of the exponential family.

\[ Y_i | \beta, \phi_i, \sim \text{Poi}(e^{\mu_i}); \mu_i = x_i' \beta + \phi_i; \]

\[ \phi = \{ \phi_i \}_{i=1}^n \sim G_n; \quad G_n | \pi_{u_1,..,u_n}, \theta = \sum_{u_1,..,u_n} \pi_{u_1,..,u_n} \delta_{\theta_{u_1}} \ldots \delta_{\theta_{u_n}} \]

\[ \pi_{u_1,..,u_n} = P \left( \sum_{k=1}^{u_1-1} p_k < F^{(i)}(\gamma_1) < \sum_{k=1}^{u_2} p_k , \ldots , \sum_{k=1}^{u_n-1} p_k < F^{(n)}(\gamma_n) < \sum_{k=1}^{u_n} p_k \right) \]

\[ \theta = (\theta_1, ..., \theta_K) \text{iid} \sim N(0, \sigma^2); \quad p_1 = V_1; \quad p_j = V_j \cdot \prod_{k<j} (1 - V_k); \quad V_j \text{i.i.d. Beta}(1, \alpha) \]

\[ \gamma = (\gamma_1, ..., \gamma_n) \sim N_n(0, \Sigma); \quad \Sigma^{-1} = \sigma^{-2}(D - \rho W). \]

Note that \( \Sigma \) above corresponds to a proper CAR distribution and \( \rho \in (0, 1) \) to ensure the existence of the inverse (Banerjee et al., 2004; Ch.3). The marginal distribution of \( G^{(i)}(\phi_i) \) is, for each \( i \), given by

\[ G^{(i)}(\phi_i) | \theta, p = \sum_{k=1}^{K} \sum_{u_1,..,u_i=k} \pi_{u_1,..,u_i=k,..,u_n} \delta_{\theta_{u_1}} \ldots \delta_{\theta_{u_i}} \]

\[ = \sum_{k=1}^{K} P \left( \sum_{l=1}^{k-1} p_k < F^{(i)}(\gamma_i) < \sum_{l=1}^{k} p_k \right) = \sum_{k=1}^{K} p_k \delta_{\theta_k} \]

The distribution doesn’t depend on \( i \), indicating that marginally all random effects follow an identical DP. Given the probabilities \( p_1, ..., p_m \) (truncate the stick-breaking process at \( m \) terms), the covariance between \( \phi_i \) and \( \phi_j \) is

\[ \text{Cov}(\phi_i, \phi_j) = \sigma_s^2 \sum_{l=1}^{m} P(u_i = u_j = l) \]

\[ = \sigma_s^2 \sum_{l=1}^{m} P \left( \sum_{k=1}^{l-1} p_k < F^{(i)}(\gamma_i) < \sum_{k=1}^{l} p_k , \sum_{k=1}^{l-1} p_k < F^{(j)}(\gamma_j) < \sum_{k=1}^{l} p_k \right) \]

\[ = \sigma_s^2 \sum_{l=1}^{m} p_l P \left( F^{(i)-1}(\sum_{k=1}^{l-1} p_k) < \gamma_i < F^{(i)-1}(\sum_{k=1}^{l} p_k) | F^{(j)-1}(\sum_{k=1}^{l-1} p_k) < \gamma_j < F^{(j)-1}(\sum_{k=1}^{l} p_k) \right), \]
where \((\gamma_i, \gamma_j)\) follows a bivariate normal distribution with covariance specified by the CAR model.

### 3.5 Computational details

Posterior inference for the ARSB and ARDP models are based upon Markov chain Monte Carlo simulations (e.g., Gelman et al., 2004). The details are outlined in the Appendix.

### 4 Simulation Example

In order to evaluate our new methods, we carried out a simulation study on a Minnesota map. 50 datasets were simulated from a Poisson distribution with the mean of each county mapped on Figure 1. The data were generated using five different \(\mu_i\) values (in model (1)) corresponding the five different shades in Figure 1. There are 47 “true boundaries” delineating the different clusters, i.e. these 47 boundaries are county borders that separate two areal units with different \(\mu_i\) values. To make the problem more interesting, we also included one county (shaded white in Figure 1) that is totally surrounded by a different cluster.

Subsequently, we fit three models to those datasets and performed statistical boundary analysis. Each of these models had the likelihood arising from 1, but differed in their specification of the spatial effects. In the first model we assigned the usual parametric CAR specification (as in (3)) to the spatial effects. The parametric CAR model does not render itself to probabilistic boundary analysis (since \(P(\phi_i = \phi_j)\) will always be zero). However, one could fit parametric CAR models and use the posterior expectation of the absolute differences of the rates, i.e. \(E(\|\eta_i - \eta_j\| \mid Data)\), where \(\eta_i = \frac{\mu_i}{E_{\gamma_i}}\) acts as a boundary difference score. Higher values will indicate spatial barriers between units \(i\) and \(j\). This approach was
used by Lu and Carlin (2006) and we call this the “LC” method.

In the second model we assigned the usual DPM model without spatial information about the map, while the third and fourth models had the ARSB and ARDP specifications respectively. The DPM, ARSB and ARDP models not only yield estimates of $\eta_i$, as in the “LC” method, but they will also deliver the posterior probabilities $P(\phi_i = \phi_j \mid i \sim j, \text{Data})$.

To assess the performance of these models, we employed the minimum predictive loss approach of Gelfand and Ghosh (1998) for model choice. Specifically, for each posterior sample $\beta^{(l)}$ and $\phi^{(l)}$, $l = 1, \ldots, L$ obtained using Markov chain Monte Carlo, we generate replicates for each data point as $y_{\text{rep},i}^{(l)} \sim \text{Poi}(\mu_i^{(l)})$, where $\mu_i^{(l)} = x_i^T \beta + \phi_i^{(l)}$. Preferred models should perform well under a decision-theoretic balanced loss function that penalizes both departure from fit and departure from smoothness as reflected by variation in replicates. We compute the departure from fit, say $G$, as $\sum_{i=1}^{n} (y_{\text{obs},i} - \mu_{\text{rep},i})$, where $\mu_{\text{rep},i} = (1/L) \sum_{l=1}^{L} y_{\text{rep},i}^{(l)}$, and the departure from smoothness, say $P$, as $\sum_{i=1}^{n} \sigma^2_{\text{rep},i}$, where $\sigma^2_{\text{rep},i} = (1/L) \sum_{l=1}^{L} (y_{\text{rep},i}^{(l)} - \mu_{\text{rep},i})^2$. This yields a model comparison score, $D = G + P$, with lower values of $D$ suggesting better model performance.

The results of the simulation example are summarized in Table 1. The promise of the models is evident from the superior performances of the ARDP and the ARSB models using all three criteria. Since we know the true boundaries in Figure 1, we can assess the performances of these approaches in detecting the true boundaries. Using direct posterior probabilities to obtain the top 47 boundaries, we find that the DPM, ARDP and the ARSB models are each able to detect about 90% of the true boundaries. The ARDP model performs slightly better than the other two, who are approximately equally good.

The LC method cannot produce these posterior probabilities. Therefore, we used the posterior expectation metric to compare its performance. We again find that the proposed ARSB and ARDP models clearly outperform the LC method. The ARDP model has almost
a 10% better detection rate, while the ARSB model excels by approximately 5%. Both ARDP and ARSB outperform the DPM model as well in terms of the posterior expectation metric. Finally, the predictive loss model choice metric D also reflects substantial evidence in favor of the ARDP and ARSB model than for the LC (i.e. parametric CAR) model and DPM model.

5 Analysis of the Minnesota P & I dataset

We illustrate our model comparison approach in the context of a Minnesota Pneumonia and Influenza diagnosis dataset. Influenza and pneumonia are major causes of illness and death. In 2005, these conditions ranked as the eighth leading cause of death in the United States and the sixth leading cause in people over 65 years of age. An active surveillance program for an influenza-like illness can help impede the spread of the infection by appropriate intervention. Boundary analysis can help identify “health barriers” separating counties that experience difference impacts of the influenza virus. Reported difference boundaries can provide information to the surveillance system to limit unnecessary entry or egress of people from the affected areas, thereby thwarting the spread of the infection. Furthermore, difference boundaries can improve coordination between neighbors and execute plans for hospital needs and antiviral or vaccine interventions.

Our dataset includes subjects older than 65 years who were enrolled in both Medicare part A and part B in December 2001. Residents of Minnesota who were 65 years of age and older and who were enrolled in the Medicare fee-for-service program as of December 31, 2001, formed our study population. This population had been identified as part of a multi-year study regarding the impact of vaccinations on elderly Minnesota residents. The Medicare Denominator file for 2001 was used to define the cohort. In addition to meeting
the age and state of residence criteria, to be eligible for inclusion in the study, the person had to be enrolled in both Medicare Part A and Medicare Part B, not be enrolled in a Medicare Advantage health plan, and not have end-stage renal disease. The denominator file also indicated the county of residence for each person.

Hospitalizations for Pneumonia and Influenza (P&I) were identified by the Medicare Provider Analysis and Review (MedPAR) short stay inpatient file for the above Minnesota residents. This annual file contains extensive patient records per hospitalization based on the date of discharge. Rates of P&I hospitalization are traditional measures of the impact of influenza virus in the elderly population. We identify the ‘boundaries’ that separate the more affected areas from the less affected areas. Here we studied the number of hospitalizations from P&I in both influenza and a shoulder period among persons at risk in each county.

Let \( Y_i \) be the observed number of hospitalizations in county \( i \), \( O_i \) be the population of county \( i \), \( E_i = \frac{\sum_{k=1}^{n} Y_k}{\sum_{k=1}^{n} O_k} O_i \) is the expected number of cases (under the assumption of no spatial variation in rates), where \( n \) is the total number of counties. We employ our four aforementioned models in Section 4 to detect boundaries on the (P&I) hospitalization map. The choropleth map of the raw data is shown in Figure 2. The high valued SMR (standard mortality ratio) counties are scattered over the map, with a clump on the southwest and some isolated regions surrounded by sparsely inhabited counties that also have lower counts. The \( D(= 50) \) most likely difference boundaries detected by each model are highlighted in Figure 3. Here about 90% of them are detected by all four models. As a specific example consider Cook and Koochiching county. The outcome variable in the former is substantially higher than its only neighbor, Lake, while Koochiching county is separated from all its neighbors due to its extremely high (P&I) SMR even after being smoothed by the model. Table 2 presents a more comprehensive “lookup table” containing the names of adjacent counties that have been ranked in decreasing order according to \( 1 - P(\phi_i = \phi_j \mid Data) \) from
the ARDP model. This table offers an easy reference for health administrators and officials
to identify the more substantial spatial health barriers in the state.

In order to compare the performance of the different models on this dataset, we again
adopt the predictive loss model choice criterion. The D scores are summarized in Table 3
which reveals that all four models perform almost equally well here in terms of the criterion
based upon departure from the ‘fit’ and the departure from the ‘smoothness’. The LC model
performs slightly better than the rest of the three and the nonspatial DPM model is slightly
inferior for this specific dataset.

6 Conclusion and Future Work

The paper presented a class of nonparametric Bayesian hierarchical model in comparison to
existing methodologies for detecting difference boundaries on maps. An advantage of the new
approach is that it permits the probabilistic estimation of an edge as a difference boundary,
and improves the percentage of true detection. A disadvantage is that the model cannot be
easily fit into any existing commercial software. We fit these models in R (www.r-project.
org), and our eventual goal is to collect these Bayesian nonparametric hierarchical spatial
models in an R package.

In Section 5, the decision to classify a boundary as a wombling boundary is arrived at by
identifying all edges \((i, j)\) such that \(P(\phi_i = \phi_j \mid Data)\) exceeds a threshold or cutoff value \(t\) or
the \(M\) boundaries with the lowest probability \(P(\phi_i = \phi_j \mid Data)\). Some more formal choices
are possible. For instance, a decision theoretic approach will treat the spatial boundary
analysis problem as one of multiple hypothesis testing. For each pair of adjacent regions,
say \(i\) and \(j\), we seek to test \(\phi_i = \phi_j\) against \(\phi_i \neq \phi_j\). This produces as many hypothesis as
there are edges, leading to a multiple hypothesis problem. Recently, several authors have
advocated the use of the false discovery rate (FDR) to adjust for multiplicities in hypothesis testing problems (see, e.g., Benjamini and Hochberg, 1995; Efron et al, 2001; Storey, 2002, 2003). Controlling the FDR is a practical and powerful approach to multiple testing and we identify this as an area of future exploration.

7 Appendix: Computational details

Posterior inference for our models are based on MCMC posterior simulations. There are two main strategies used. The first avoids computing parameters characterizing $G$ by marginalizing it out and relying on the Polya urn scheme of Blackwell and MacQueen (1973). A limitation of this approach is that it is only applicable when the prior can be characterized by a generalized Polyn urn mechanism. Ishwaran and James (2001) proposed the blocked Gibbs Sampler that directly sampled from the posterior of the random measure, avoiding the marginalization over $G$. We use the blocked Gibbs Sampler with some Metropolis-Hasting steps nested in it to update all random parameters in our model. We truncate the infinite sum in $G$ by the first $m$ terms. We only provide details for the ARDP model. That for the ARSB model is similar (and even simpler), while algorithms for the DPM model may be found in Blackwell and MacQueen (1973).

7.1 Posterior computation for the ARDP model

We place a flat prior on parameter $\beta$ and reparameterize the variance parameters with its inverse, $\tau_s = \sigma_s^{-2}$, $\tau_\gamma = \sigma_\gamma^{-2}$, then place a conjugate gamma prior of the precision parameters $\tau_s$ and $\tau_\gamma$. The likelihood of the model is expressed as

$$L = \prod_{i=1}^{n} PoI(Y_i | x_i'\beta + \phi_i)$$

(6)
The posterior density given the data \( Y = \{ Y_i \} \) is proportional to the likelihood multiplied by all the prior distributions:

\[
L(Y_i | \beta, \{ \phi_i \}) \sim p(\beta)p(\phi | \tau_s)p(\gamma)p(V)p(\tau_s)p(\tau_\gamma)
\]

Note that \( \phi_i = \theta_{u_i} \) and we updated \( \phi_i \) by updating \( \theta \) and \( u_i \). The MCMC algorithm proceeds as follows.

Step 1: update \( \beta | \theta; \gamma, V, \tau_s, \tau_\gamma \): The full conditional distribution only depends on the likelihood due to the flat prior. Sample candidate \( \beta^* \) from \( N(\beta, 0.05I) \), then accept the candidate \( \beta^* \) with probability

\[
\min \left\{ 1, \frac{\exp(\sum_{i=1}^n (\gamma - \bar{y}_i + \gamma))}{\exp(\sum_{i=1}^n (\gamma - \bar{y}_i + \gamma))} \right\}
\]

Step 2: update \( \theta_j | \beta, \gamma, V, \tau_s, \tau_\gamma \): Sample candidate \( \theta^*_j \) from \( N(\theta_j, 0.05) \), then accept the candidate \( \theta^*_j \) with probability

\[
\min \left\{ 1, \frac{\exp(\sum_{i:u_i=j} (\gamma - \bar{y}_i + \gamma))}{\exp(\sum_{i:u_i=j} (\gamma - \bar{y}_i + \gamma))} \right\}
\]

Step 3: update \( \gamma_i | \beta, \theta, V, \tau_s, \tau_\gamma \): Sample candidate \( \gamma^* \) from \( N(\gamma_i, 0.01) \), compute the corresponding candidate \( u^* \) through \( u^*_i = \sum_{j=1}^n j I(\Sigma_{k=1}^{j-1} p_k < F(\gamma_i) < \Sigma_{k=1}^{j} p_k) \), then accept the candidate \( \gamma^* \) with probability

\[
\min \left\{ 1, \frac{\exp\left(-\frac{1}{2}\gamma^*\Sigma_2^{-1}\gamma^*\right) \exp(-\bar{y}_i + \gamma + \gamma)}{\exp\left(-\frac{1}{2}\gamma\Sigma_2^{-1}\gamma\right) \exp(-\bar{y}_i + \gamma + \gamma)} \right\}
\]

Step 4: update \( V | \beta, \theta, \gamma, \tau_s, \tau_\gamma \): Sample candidate \( V^* \) from \( N(V, 0.01I_m) \), compute the cor-
responding $\mathbf{p}^*$ and $\mathbf{u}^*$. Accept the candidate $V^*$ with probability

$$
\min \left\{ 1, \frac{\prod_{k=1}^{m}(1-V_k^*)^{\alpha-1}}{\prod_{k=1}^{m}(1-V_k)^{\alpha-1}} \prod_{i=1}^{n} \exp(-E_i \exp(x'_i \beta + \theta u^*_i) + y_i(x'_i \beta + \theta u^*_i)) \right\}
$$

Step 5: update $\tau_s | \beta, \theta, \gamma, V, \tau_\gamma$: Sample from the conjugate gamma posterior distribution

$$
\Gamma \left( \frac{n}{2} + c, \frac{\sum_{i=1}^{n} \phi_i^2}{2} + d \right).
$$

Step 6: update $\tau_\gamma | \beta, \theta, \gamma, V, \tau_s$: Sample from the conjugate gamma posterior distribution

$$
\Gamma \left( \frac{n}{2} + c, \frac{\sum_{i=1}^{n} \gamma(D - \rho W) \gamma_i}{2} + d \right).
$$

References


Blackwell, D. and MacQueen, J.B. (1973) Ferguson distributions via Polya urn schemes. 


Table 1: Agreement of the true wombling boundaries with those produced by LC, ARDP and ARSB based on $P(\phi_i = \phi_j)$, $E(\|\eta_i - \eta_j\| \mid Y)$ and the posterior predictive model choice metric in the simulation study

<table>
<thead>
<tr>
<th></th>
<th>Agreement by $P(\phi_i = \phi_j)$</th>
<th>Agreement by $E(|\eta_i - \eta_j| \mid Y)$</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>-</td>
<td>78.7%</td>
<td>6379.88</td>
</tr>
<tr>
<td>DPM</td>
<td>89.3%</td>
<td>82.2%</td>
<td>6450.16</td>
</tr>
<tr>
<td>ARDP</td>
<td>91.4%</td>
<td>88.3%</td>
<td>5796.01</td>
</tr>
<tr>
<td>ARSB</td>
<td>89.1%</td>
<td>83.3%</td>
<td>5920.32</td>
</tr>
</tbody>
</table>

Figure 1: A map of the simulated data with the grey-scales showing the six different clusters, each having its own mean. There are 47 boundary segments that separate regions with different means (shades)
Table 2: Names of adjacent counties that have significant boundary effects from the ARDP model. The numbers in the first column are the ranks according to their $P(\phi_i = \phi_j)$

<table>
<thead>
<tr>
<th>Rank</th>
<th>County 1</th>
<th>County 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Beltrami, Koochiching</td>
<td>26 Koochiching, Lake of the Woods</td>
</tr>
<tr>
<td>2</td>
<td>Cass, Wadena</td>
<td>27 Isanti, Mille Lacs</td>
</tr>
<tr>
<td>3</td>
<td>Douglas, Pope</td>
<td>28 Chippewa, Renville</td>
</tr>
<tr>
<td>4</td>
<td>Freeborn, Steele</td>
<td>29 Murray, Pipestone</td>
</tr>
<tr>
<td>5</td>
<td>Goodhue, Olmsted</td>
<td>30 Becker, Mahnomen</td>
</tr>
<tr>
<td>6</td>
<td>Itasca, Koochiching</td>
<td>31 Goodhue, Olmsted</td>
</tr>
<tr>
<td>7</td>
<td>Kandiyohi, Pope</td>
<td>32 Clearwater, Mahnomen</td>
</tr>
<tr>
<td>8</td>
<td>Koochiching, St. Louis</td>
<td>33 Traverse, Wilkin</td>
</tr>
<tr>
<td>9</td>
<td>Pope, Stearns Murray, Redwood</td>
<td>34 Chisago, Isanti</td>
</tr>
<tr>
<td>10</td>
<td>Anoka, Isanti</td>
<td>35 Redwood, Yellow Medicine</td>
</tr>
<tr>
<td>11</td>
<td>Dakota, Goodhue</td>
<td>36 Freeborn, Steele</td>
</tr>
<tr>
<td>12</td>
<td>Lincoln, Pipestone</td>
<td>37 Goodhue, Wabasha</td>
</tr>
<tr>
<td>13</td>
<td>Murry, Redwood</td>
<td>38 Pope, Swift</td>
</tr>
<tr>
<td>14</td>
<td>Steele, Waseca</td>
<td>39 Cass, Morrison</td>
</tr>
<tr>
<td>15</td>
<td>Renville, Yellow Medicine</td>
<td>40 Fillmore, Olmsted</td>
</tr>
<tr>
<td>16</td>
<td>Cottonwood, Murray</td>
<td>41 Hubbard, Wadena</td>
</tr>
<tr>
<td>17</td>
<td>Jackson, Martin</td>
<td>42 Douglas, Grant</td>
</tr>
<tr>
<td>18</td>
<td>Kandiyohi, Swift</td>
<td>43 Mahnomen, Norman</td>
</tr>
<tr>
<td>19</td>
<td>Pope, Stevens</td>
<td>44 Grant, Wilkin</td>
</tr>
<tr>
<td>20</td>
<td>Todd, Wadena</td>
<td>45 Mahnomen, Polk</td>
</tr>
<tr>
<td>21</td>
<td>Lyon, Redwood</td>
<td>46 Jackson, Nobles</td>
</tr>
<tr>
<td>22</td>
<td>Murray, Nobles</td>
<td>47 Morrison, Todd</td>
</tr>
<tr>
<td>23</td>
<td>Isanti, Sherburne</td>
<td>48 Dodge, Olmsted</td>
</tr>
<tr>
<td>24</td>
<td>Otter Tail, Todd</td>
<td>49 Big Stone, Traverse</td>
</tr>
<tr>
<td>25</td>
<td>Clay, Otter Tail</td>
<td>50 Morrison, Stearns</td>
</tr>
</tbody>
</table>

Table 3: Predictive loss criterion under all four models for Minnesota (P&I) dataset. G is a goodness-of-fit term while P is a penalty term which penalized departure from “smoothness”.

<table>
<thead>
<tr>
<th>Model</th>
<th>G</th>
<th>P</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>0.47</td>
<td>2.81</td>
<td>3.29</td>
</tr>
<tr>
<td>DPM</td>
<td>0.82</td>
<td>2.76</td>
<td>3.58</td>
</tr>
<tr>
<td>ARDP</td>
<td>0.76</td>
<td>2.73</td>
<td>3.49</td>
</tr>
<tr>
<td>ARSB</td>
<td>0.64</td>
<td>2.77</td>
<td>3.41</td>
</tr>
</tbody>
</table>
Figure 2: Choropleth map of the SMR in MN (P&I) dataset

Figure 3: Difference boundaries detected by various models in the Minnesota (P&I) dataset.